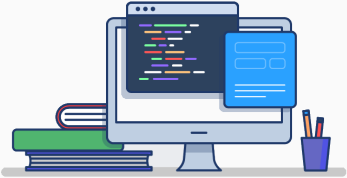
Resignation Calls and Dismissals of Ministers in Latin America

Data Gathering using NLP and Machine Learning

Bastián González-Bustamante
University of Oxford
✉ bastian.gonzalezbustamante@politics.ox.ac.uk

Presentation at the Congress of the Latin American Studies Association
May 5-8, 2022

Table of Contents



Introduction

Theoretical Gap

The literature on coalitions in presidential systems, ministerial reshuffles, and ministerial recruitment is diverse. However, there remains a lack of clarity about the relationship and incentives between presidents and cabinet members.

For example, **when is a president willing to remove a cabinet member, and when is she willing to protect him in the face of scandals, policy failures or other challenges?**

The answer lies in a number of interrelated factors, such as chains of delegation and incentives in the principal-agent relationship between the president and her ministers or between the president and the electorate (Elgie, 2020).

Resignation Calls as Proxies

Other elements can shape the presidential decision, such as different **stochastic events** that tend to operate as random shocks affecting government stability (Chiba et al., 2015; Fortunato and Loftis, 2018), or **moral hazard** and **agency problems** that can affect cabinet performance (Chaisty et al., 2018; Martínez-Gallardo and Schleiter, 2015). Indeed, the literature has identified that protests, economic crises, scandals of different sorts, among other stochastic events, do affect cabinet stability (Camerlo and Pérez-Liñán, 2015; Martínez-Gallardo, 2014).

However, as Berlinski et al. (2010) point out, it is **complex and empirically costly** to assess all possible random shocks. Therefore, cues such as **calls for ministers to resign** may be empirically efficient indicators. This information, nevertheless, is scarce in several countries and non-existent in Latin America.

NLP for Data Collection

This working paper presents a new data set created using **Natural Language Processing (NLP)** and **machine learning techniques**. The data set contains detailed information on the cabinet turnover in 12 Latin American democracies from the time of redemocratisation between the 1970s and 1980s, depending on the case, up until the end of the latest presidential terms.

To this end, we used a number of sources of public information and press reports that were digitised with data mining algorithms. Using machine learning models, we were able to identify ministerial resignation calls during the period.

This data is indeed **completely novel for Latin American presidentialisms**.

Data Collection and Creation of the Data Set

First Raw Data Set

Data collection is based on a three-stage procedure. First, a **raw data set of ministers and portfolios** is compiled according to a number of relevant variables (*i.e.*, country, name of the minister, gender, party affiliation, appointment date, etc.)

This raw data set is compiled from a revision of official sources, recognised press, information available in the libraries of Congress of each country, and data from the **Latin America Weekly Report (LAWR)**.

These weekly reports of between 12 and 16 pages long present a compendium of relevant news on each country in the region. On average, they constitute approximately 600 pages per year until 2002 and 800 pages since then. These reports are a significant source of aggregate information at a regional level.

First Raw Data Set

Second, a **novel raw data set of weekly ministerial resignation calls** is generated from a revision of media reports from LAWR. For this, the specific mentions of cabinet members are identified, allowing us to codify the variable of resignation calls. To this end, LAWR archives were compiled in three different batches of data.

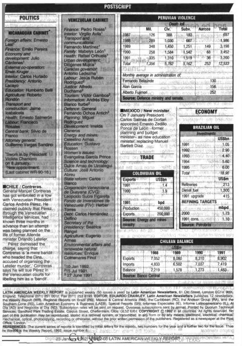
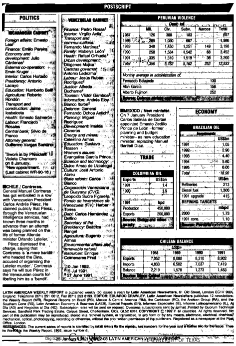
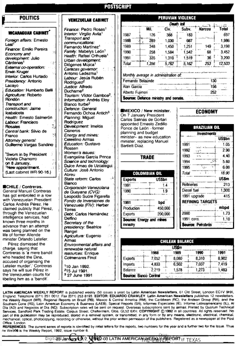
These files are stored in a private repository with a version control system on GitHub whose access is controlled by Two-Factor Authentication (2FA). In addition, it is backed up on Hierarchical File Server (HFS) for recovery in case of unforeseen incidents on the University of Oxford hub connected to Code42 Cloud Backup GUI-based, allowing version control, restoration, and automatically scheduled backups, and 256-bit Advanced Encryption Standard (AES).

Detail of the Batches of the LAWR Archives

Batch	Archives	Years	Format	Volume
Batch-01	LAWR-1975 to LAWR-1979 Including 1997	5	CSV UTF-8	9.56 MB
Batch-02	LAWR-1980 to LAWR-1998 Excluding 1997 Including the first half of 2003	18.5	PNG images	38.5 GB
Batch-03	LAWR-1999 to LAWR-2021 Excluding the first half of 2003	22.5	PDF files	341 MB

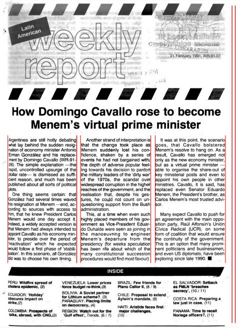
Note: Revision of 12 Latin American countries thanks to St Hilda's College Muriel Wise Fund and subscriptions of the Bodleian Libraries at the University of Oxford.

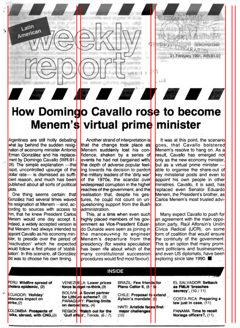
Binarisation Algorithms

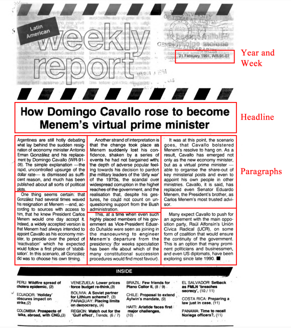




Note: Images correspond to a LAWR page before binarisation and after the application of techniques of Sauvola and Pietikäinen (2000) and Su et al. (2010).

Tesseract OCR Algorithm

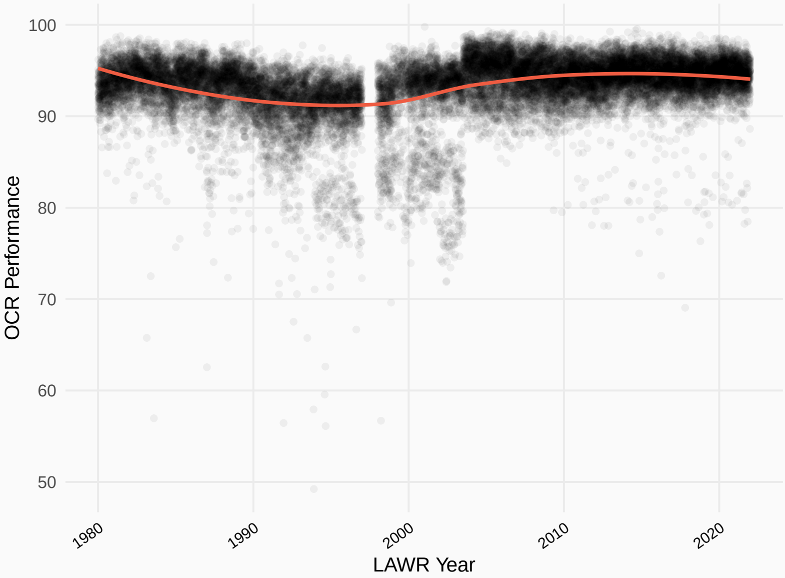






Tab-Stop Lines — Column Layout — Segmented Blocks

Accuracy of the OCR Algorithm



Data Structure

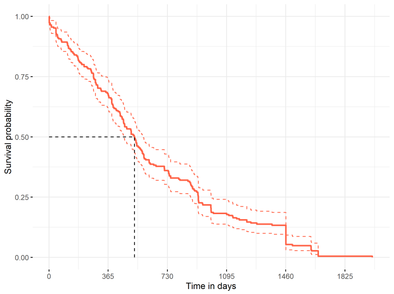
ℹ OCR application on **41 years** of files with a total of **28,090 pages**.

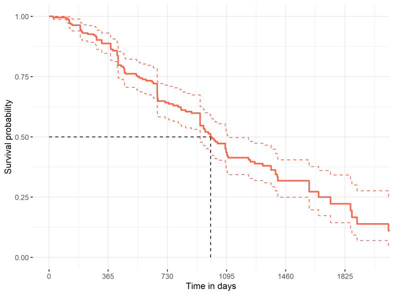
ℹ We identified **18,198 mentions** of ministers and cabinets members (incorporating the first batch this increases to **19,925** and the coverage to **46 years**).

We elaborate a weekly data set with resignation calls to merge with the first raw data and other sources such as approval, party systems variables, and macroeconomic indicators.

Therefore, we encode the data as **time-dependent** covariates accounting for changes over time.

Examples of Kaplan-Meier Survival Estimations



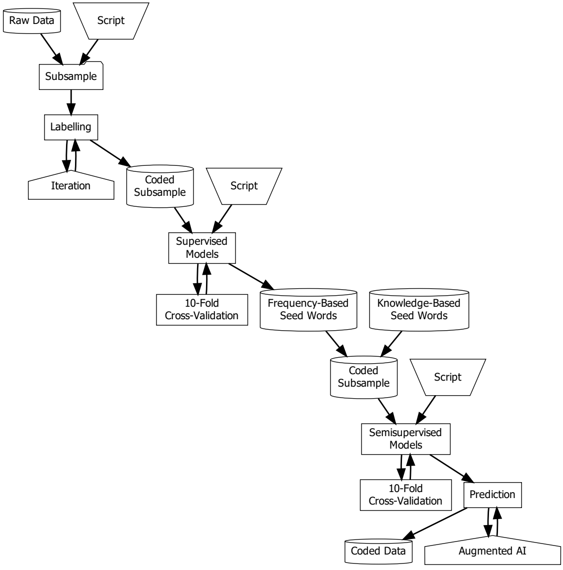


Note: The plots show Argentina and Chile with a 95% confidence interval and median survival.

How do we distinguish between mentions and calls?

Machine Learning Models to Classify Resignation Calls

Machine Learning Algorithms Training Pipeline



Labelling for Training Supervised Models

Random subsample of **1,000 LAWR press releases** considering cabinet mentions. Assets pool of 3,000 observations (three hand-coded labels per unique report).

We use the online labelling platform Labelbox (2021), which subsequently allows all the information to be exported in JavaScript Object Notation (JSON) format for processing.

The labelling was conducted by **six human coders**, involving 41 hours and 55 minutes of actual work on the platform using Dewan and Dowding's (2005) categorisation: personal scandals, financial scandals, policy failures, internal disagreements, or other controversies.

Performance of the Labelling Process

Coder	Labels	$M(T)$ Label	$\sum$ Time	Consensus
1	966	39s	10h 25m	94.05
2	718	1m 15s	15h 0m	94.84
3	704	38s	7h 29m	94.82
4	318	1m 4s	5h 38m	95.47
5	263	17s	1h 14m	99.26
6	101	1m 17s	2h 9m	88.50

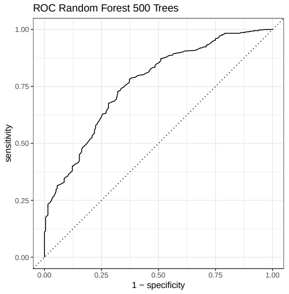
Note: Krippendorff's α (bootstrap of 1,000 iterations) $= 0.870$ and $\mathrm{CI}_{95\%}$: 0.844 to 0.900.

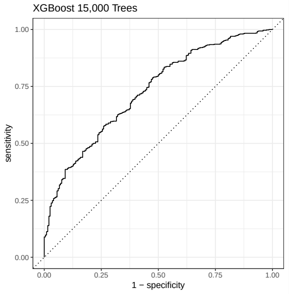
Considering a binary classification of resignation calls $\alpha = 0.895$ and $\mathrm{CI}_{95\%}$: 0.868 to 0.922.

Performance of the Supervised Classifiers

Model	Snowball 500 Tokens		
	Accuracy	Precision	Recall
Naïve Bayes	0.714	0.803	0.825
Kernel Linear SVM	0.658	0.802	0.728
Kernel Gaussian RBF SVM	0.756	0.756	0.999
Random Forest (100 trees)	0.761	0.760	0.998
Random Forest (500 trees)	0.761	0.760	0.999
XGBoost	0.784	0.811	0.932
Model	**Stem. SMART 100 Tokens**		
	Accuracy	Precision	Recall
Naïve Bayes	0.725	0.786	0.876
Kernel Linear SVM	0.735	0.785	0.894
Kernel Gaussian RBF SVM	0.766	0.766	0.993
Random Forest (100 trees)	0.775	0.773	0.995
Random Forest (500 trees)	0.771	0.770	0.995
XGBoost	0.771	0.809	0.912

ROC Curve for Random Forest and XGBoost





Performance of the Semisupervised Classifiers

Model	Knowledge-Based Seeds		
	Accuracy	Precision	Recall
Naïve Bayes	0.713	0.734	0.948
Kernel Linear SVM	0.698	0.741	0.900
Kernel Gaussian RBF SVM	0.726	0.728	0.995
Random Forest (100 trees)	0.743	0.742	0.991
Random Forest (500 trees)	0.743	0.740	0.997
XGBoost	0.726	0.762	0.905
Model	**Frequency-Based Seeds**		
	Accuracy	Precision	Recall
Naïve Bayes	0.758	0.805	0.918
Kernel Linear SVM	0.771	0.808	0.932
Kernel Gaussian RBF SVM	0.793	0.794	0.997
Random Forest (100 trees)	0.799	0.797	0.999
Random Forest (500 trees)	0.801	0.799	0.999
XGBoost	0.794	0.819	0.948

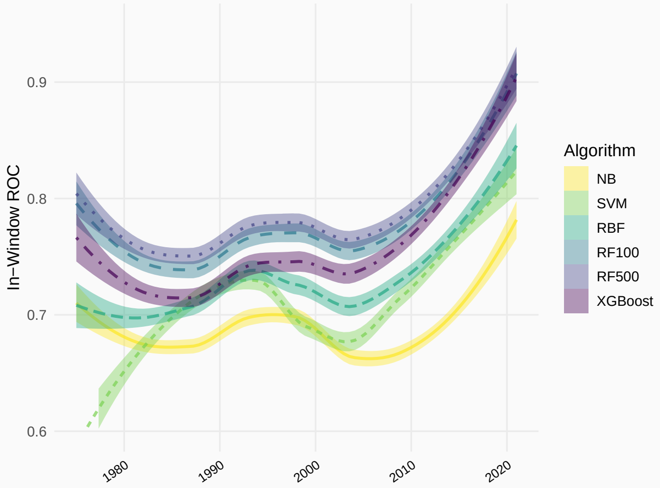
Seed Words and Fixed Rolling Windows

If we consider both seed words sets, the level of agreement between machine and human labelling, considering our subsample coded on Labelbox (2021), is 69.2% for machine and two human coders.

We follow a strategy similar to that used by Greene et al. (2019): **Five-year fixed rolling window** between 1975 and 2021 to train our models and predict resignation calls.

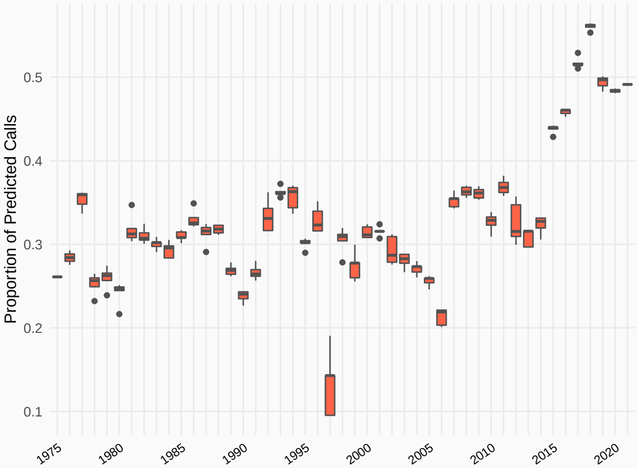
Therefore, the first window covers the period from 1975 to 1979, the second between 1976 and 1980, and successively until the forty-third window from 2017 to 2021.

Performance of the Semisupervised Models (1975-2021)



Note: Indicators 10-fold cross-validated in the five-year rolling window.

Predicted Ministerial Resignation Calls (1975-2021)



Note: Predictions proportion over the mentions carried out with the ensemble semisupervised Random Forest (500 trees) in the five-year rolling window.

Concluding Remarks

Concluding Remarks

This working paper presented a pioneering application of OCR procedures and machine learning models to build a data set with **novel information** on cabinet members' calls for resignation in **12 Latin American democracies**. This can help study coalition dynamics in presidential systems, cabinet management and stability, political recruitment processes at the ministerial level, etc.

This is still a **work in progress**. Further work on distinguishing resignation calls in more detail is necessary. In addition, the data set needs to be contrasted with other similar sets.

Our data set permits the outlining of empirical strategies with and without causal identification strategies to estimate different types of nonparametric and parametric econometric models, as well as competing risks and survival analysis approaches.

Acknowledgements and Data Sharing

Acknowledgements

This research has been funded by the Chilean National Agency for Research and Development (ANID/PFCHA/72200340), the St Hilda's College Muriel Wise Fund, and the Society for Latin American Studies Postgraduate & Postdoctoral Research Support Award.

My thanks to Petra Schleiter for her substantial comments and support. I also thank Carla Cisternas and Diego Aguilar for their help during the file review and validation process.

Finally, thanks to the human coders who conducted the labelling process to train the supervised models.

Preservation and Data Sharing

The data set will be deposited in the digital repository Oxford Research Archive for Data (ORA-Data) in CSV UTF-8 format with its codebook and standardised metadata. The data set will remain under embargo until October 2023 (to be confirmed).

It will then be available for reuse under an open-access licence that allows sharing and adapting the material without additional restrictions as long as appropriate acknowledgement is given.

Further technical details on storage, formats and replicability are available in the Data Management Plan (DMP ID 85349).

References and Contact Information

References I

Berlinski, S., Dewan, T., and Dowding, K. (2010). The Impact of Individual and Collective Performance on Ministerial Tenure. *The Journal of Politics*, 72(2):559–571.

Camerlo, M. and Pérez-Liñán, A. (2015). Minister Turnover, Critical Events, and the Electoral Calendar in Presidential Democracies. *The Journal of Politics*, 77(3):608–619.

Chaisty, P., Cheeseman, N., and Power, T. J. (2018). *Coalitional Presidentialism in Comparative Perspective: Minority Presidents in Multiparty Systems*. Oxford University Press, Oxford.

Chiba, D., Martin, L. W., and Stevenson, R. T. (2015). A Copula Approach to the Problem of Selection Bias in Models of Government Survival. *Political Analysis*, 23(1):42–58.

Dewan, T. and Dowding, K. (2005). The Corrective Effect of Ministerial Resignations on Government Popularity. *American Journal of Political Science*, 49(1):46–56.

Elgie, R. (2020). Presidents and Cabinets. In Andeweg, R. B., Elgie, R., Helms, L., Kaarbo, J., and Müller-Rommel, F., editors, *The Oxford Handbook of Political Executives*. Oxford University Press, Oxford.

References II

Fortunato, D. and Loftis, M. W. (2018). Cabinet Durability and Fiscal Discipline. *American Political Science Review*, 112(4):939–953.

Greene, K. T., Park, B., and Colaresi, M. (2019). Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects. *Political Analysis*, 27(2):223–230.

Labelbox (2021). Labelbox. Online. Available at https://labelbox.com.

Martínez-Gallardo, C. (2014). Designing Cabinet: Presidential Politics and Ministerial Instability. *Journal of Politics in Latin America*, 6(2):3–38.

Martínez-Gallardo, C. and Schleiter, P. (2015). Choosing Whom to Trust. Agency Risks and Cabinet Partisanship in Presidential Democracies. *Comparative Politics Studies*, 48(2):231–264.

Sauvola, J. and Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236.

Su, B., Lu, S., and Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. In *DAS '10: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, Boston. ACM Press.

Contact Information

Bastián González-Bustamante

DPhil (PhD) Researcher

Department of Politics and International Relations
& St Hilda's College
University of Oxford
St Hilda's College, Cowley Place, Oxford OX4 1DY
bastian.gonzalezbustamante@politics.ox.ac.uk
https://bgonzalezbustamante.com

Presentation compiled with LaTeX and a lot of
Download the latest version from GitHub

Thank you very much!